DOCUMENT RESUME

ED 286 922 TM 870 557

AUTHOR Ingebo, George S.
TITLE Riding the Rasch Tiger. Part 2: Implications for District Testing Programs (Where the Rubber Meets the Road).
PUB DATE Apr 87
NOTE 45p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Washington, DC, April 21-23, 1987). For parts 1 and 3, see TM 870 556 and TM 870 558.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143) -- Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Rating; Achievement Tests; Basic Skills; Difficulty Level; Elementary Secondary Education; *Item Analysis; *Item Banks; *Latent Trait Theory; Local Norms; Mathematical Models; Measurement Objectives; Measurement Techniques; School Districts; Tables (Data); Test Construction; Testing Problems; Test Items
IDENTIFIERS *Calibration; Curriculum Related Testing; Equal Interval Scoring; Northwest Evaluation Association; Portland Basic Skills Achievement Levels Testing; *Rasch Model

ABSTRACT

Greater knowledge about the practical application of Rasch technology can help in avoiding misapplications and confusions in testing programs. Equal interval curriculum-based scaling makes possible the following improvements in measuring basic skills achievement by enabling testing programs to: (1) individualize the difficulty level of group-administered tests; (2) modify tests longitudinally; (3) provide high-level test security; (4) add locally developed items; (5) use equal interval measures to assess progress and develop local norms; (6) make comparisons using standardized differences; (7) construct score conversion tables; (8) provide district testing information; (9) facilitate program evaluation; and (10) provide a defensible basis for cross-grade groupings of students with comparable basic skills achievement. Appendix 1 presents longitudinal calibration analyses tables examining interitem calibration relationships across grades. Appendix 2 is a guide for making achievement comparisons, in the format of a series of questions, each followed by one or more explanations of how the data can answer that question. Comparisons using Rasch units (RITs), Portland (Oregon) standard scores (P-scores), RIT growth indices, RIT rate indices, and RIT composite scores are considered. The final section of this guide offers a procedure for reporting test results to different user groups. (LPG)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED286922

RIDING THE RASCH TIGER

PART 2: IMPLICATIONS FOR DISTRICT TESTING PROGRAMS
(Where the Rubber Meets the Road)

George S. Ingebo

Portland Public Schools (Retired)
501 N. Dixon Street
Portland, Oregon 97208

Presented at
The National Association of Test Directors Symposium
The National Council on Measurement in Education
Annual Conference
April 1987
Washington, D.C.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. Ingebo

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM 870 557

BEST COPY AVAILABLE

RIDING THE RASCH TIGER: IMPLICATIONS FOR DISTRICT TESTING PROGRAMS

My purpose is to get across two ideas that can make the rest of your professional lives busier and much more enjoyable when it comes to acquiring basic skills achievement information. Thanks to the Rasch's tigerish nature, we can: 1) think in terms of a testing system, and 2) enjoy the flexibility inherent in using calibrated item banks to do real test system engineering. We will list 10 important aspects of test adaptation to basic skills achievement that cannot be offered by the old line traditional test development methods - it's an exciting ride.

In the first National Association of Test Directors (1986) presentation of the Rasch Tiger Series, Taming The Rasch Tiger, discussants pointed out some of the important things to know about practical application of Rasch technology, some misuses of the technology that will put results in doubt, and some of the irrelevant arguments that distract attention from the reality of measurement.

Thomas' (1986) statements picked up and underlined by Coffman (1986), regarding stability of items calibrations warns a test developer to keep his eye on the ball -- points out some things to know. Item calibrations stability is one of the most important and it must be experimentally determined over the full range of the curriculum bank and full range of instructional program for which tests are being constructed. A caution here is that one determine and then validate calibrations. Then pragmatically compute how much calibrations vary by determining actual change over instructional groups rather than using the standard error of calibration provided by the usual Rasch computer programs. The Appendix #1 table, first programmed by Dr. Forster to examine levels test items stability in 1981 following experimentation presented at Washington Educational Research Association of the 1980 conference, shows how stability of an item in the Northwest Evaluation Association Item Bank is established through administration of five or six large samples of students in five or six different grades taking the same test because it is at their achievement level and they are personally involved as part of an ongoing evaluation system. Coffman also called our attention to misuses of the Rasch, or any probablistic model for that matter, when logic is abandoned and computer programs are idolized. Rasch technology has been thoroughly researched and proven effective in providing equal interval measurement in arithmetic, reading, and language usage basic skills only. Basic skills are essentially man-made rules and to some extent are sequential in nature. In order to determine the feasibility of scaling a curriculum like arithmetic as a global survey test, computation, problem application, and concept items were Rashed separately and together. The conclusion was that an arithmetic curriculum scale was defensible. Forbes and Ingebo (1975). This provides the bases for curriculum scales that permit obtaining equal interval measures rather than just population based scores.

Calibrating some aspects of the curriculum seems dubious and for other aspects very doubtful. There is a sampling problem as well as a lot of slippage in testing spelling, and vocabulary seems even less likely to scale nationwide at the upper end. It's an example of misuse of the Rasch. In contrast to basic skills sciences, including social science, are largely efforts to discover and quantify relationships that exist in nature. Although data from a test over aspects of "oxidation and reduction" or a test over "the Westward Movement" can be run through a computer, to my knowledge the data/logic confluence for measurement scaling to the science or social science curriculum hasn't been written. It seems quite evident that a sort of curricular phalansterianism has frustrated attempts to define such a measurement scale. To generalize, it's best to be suspicious of "misuse" of a mathematical formula applied to human responses until it has been well tested through application in a genuine setting. A problem related to misuse is publishing information that the model won't "work." It's embarrassing to see a naive argument of that nature when one has eight years of very successful, practical, data from using Rasch technology.

The Rasch model has attracted more irrelevant arguments and more data free response than any test improvement techniques that I have seen surfaced since I started building items analyzed achievement tests 38 years ago. Often the word "assumptions" is used without reference to any basic reality of measuring. Actually the only assumption to be met in making use of a test construction technology is that students who know more will answer questions correctly more often than students who know less. Using the Rasch technology requires ingenuity, hard work and sensitivity to data quality on the test developers part as well as knowing what it can and cannot do for you. Another misunderstood verbalization has to do with "items fitting the model." The real focus of attention should be on the "item fitting the curriculum." Item "independence" is another verbal surprise. Most knowledge and skills in the curriculum area had better have something in common with each other. "Item interaction" with others that provide cues is a test construction danger.

I have a simplistic explanation for much of the overextended tangential rationalizing that is broadcast on various self-enhancing wave-lengths - if you bear with me in recalling an old shaggy dog problem. Once when a trio of deer hunters were snowed in at Harlotown, Montana, they checked into a hotel and each paid $10.00 for a suite of rooms to the bellboy who was relieving the desk clerk at the time. When the regular desk clerk returned he told the bellboy that the rates for three people was $25.00 and gave him $5.00 to return to the overcharged hunters. On the way up in the elevator the bellboy recognized the difficulty of dividing $5.00 three ways. He solved that problem by pocketing $2.00 and returning $1.00 to each of the hunters. The dilemma of the disappearing dollar surfaces when you consider the fact that: 1) after the refund each hunter paid $9.00 which comes to $27.00, 2) the bellboy has the other $2.00, 3) in accounting for the dollars the $27.00 the hunters paid and the $2.00 the bellboy kept adds up to only $29.00.

Now the original payment was $30.00, but 3 x 9 = $27.00 and with bellboy's $2.00, only $29.00 are accounted for. But dollars shouldn't evaporate.

Utter confusion because $2.00 and $27.00 fall short of the original $30.00 that was paid in. Why? When one uses parts of different problems to produce an answer, it can be perplexing. Actually; 30 - 3 = 27, 25 + 3 + 2 = 30, 27 + 3 = 30, 3 + 3 x 9 = 30, but if you take one part of the third problem ($27.00) and one part of the second problem ($2.00) as sufficient information, you don't solve either problem.

Human beings tend to cast about for explanations when placed in a novel somewhat complex situation and when a strong personal bias is in the background they have a tendency to mix components of two or more different problems as shown in the "case of the missing money" to come up with strange conclusions and unwarranted generalizations. If one mixes problems commonly subsumed under the heading of "guessing" when item ambiguity is evidenced in numbers or when one mixes in those mathematical abstractions called "discrimination" with other data designed to free measurements from populations one can expect to come up with confused results.

Another confusion in thinking comes under the name of "vertical equating." The literature contains a few well done attempts to do something that should not be expected to succeed at all. Tests designed for a spread of ability found in one grade are too inappropriate for too many students to scale with each other. Actually, the best chance would be to test an intact, clear group of students a year later with the second test, but why bother. Just build short pre-equated tests using the callibrated bank.

When a clear discontinuity of testing procedures of the magnitude introduced by using Rasch technology is proposed it is easy for some whose favorite indoor sport is designing models to give too little attention to the realities of student/curriculum/teacher/instruction quadrangles and talk as though their names for various statistics mean something - as though names are sufficient for informative arguments in making data free assertions. Thanks again to Forbes, Thomas, and Coffman (1986) and others in Taming the Rasch Tiger for putting the use of Rasch technology in perspective for potential users.

At this time our research and nine years of implementation has used only the Northwest Evaluation Association item banks so the word bank will reference to the NWEA banks. Since we have empirically established the integrity of item calibrations in the banks (see Appendix I), let's consider the other half of testing fit with the curriculum. There is no way that I know of to statistically correct for item mismatch with curriculum specifications. Goals and objectives that items were written to test need to be clearly stated and are used to select out of the 3,000 or more questions in N.W.E.A. item banks. However, selecting and accepting an item for a test requires judgement calls by curriculum and instruction professionals in the educational organization making use of the test information (Wolmut 1975). Using a test with items that aren't in themselves appropriate for measuring students in a district is counterproductive.

To assemble a test for review and revision by curriculum folk:

1. The test blueprint must be specific in the nature of the information to be obtained about what students know or are able to do.

   1.1 Goal statements and/or objectives must be clear.
   1.2 Balance -- number of items in each goal/objective stated.

2. Each item must relate directly to the curriculum and instructional program (not necessarily methodology) of the district.

3. Difficulty level of each test stated in ranges of item difficulty must be established.

Once we have our marching orders from the curriculum and/or instructional professionals in terms of the test blueprints -- formats that are balanced in number of items measuring each identified content area -- we can proceed to articulate the tests necessary and sufficient in a comprehensive testing system. Of course these will be returned to Curriculum and Instruction for review and revision as provided for by district policy. But at this time in our professional lives, the careful background work with the Rasch pays off in building useful measuring instruments.

The main use of Rasch technology is establishing the level, often referred to as difficulty, ("difficulty" refers to difficult for somebody or some group but the item "level" refers to the items position on the curriculum scale), of every item in a bank compared to the curriculum and to the level of every other item in the bank. This was done for all three item banks by carefully testing about a quarter of a million students for each bank. Linking field tests was the most important contribution of the Rasch. Forster, F. and Ingebo, G. (1976).

What the banks provide us is <u>flexibility</u> in test construction - flexibility never before enjoyed by basic skills test developers. Each item in the bank is a separate unit with <u>known properties</u> so items can be combined with some other items to format a test that will meet required specifications, like bricks used to build a wall. Contrast ordering a standard rectangular prefabricated wall with ordering a truck load of bricks and building a wall shaped in what ever way it is best to serve your purposes. Equal interval curriculum based scaling and the flexibility it makes possible are the keys to the most important development of basic skills testing in the 20th century, including machine scoring, because it makes possible the following 10 changes in measuring achievement -- makes possible a testing system publishers of conventional tests cannot duplicate through traditional methods.

1. <u>Individualizes</u> group administered tests according to each student's achievement record.

2. <u>Modifies</u> tests by deleting, substituting, or adding items and maintaining longitudinal records.

3. Provides high level test security for regular group testing and at maximum for competency/standards testing which needs to withstand legal challenges.

4. Adds locally developed items through the fixed parameter method. These can be included in an item bank for a district.

5. Provides equal interval measures as the basis for assessing progress. Means and standard deviations for District, schools, classes, etc. are more defensible than those based on population data only.

6. Allows use of standardized differences between measures for describing small or large groups because data is on equal interval scales.

7. Allows score conversion table to be made for any type of test before it is administered to any of the students.

8. Provides district testing system information about all school basic skills programs, including federally funded programs, that need basic skills achievement information.

9. Facilitates program evaluation through groups with number of students of class size and greater. Control grouping and regression to the mean problems managed through research design that is made possible by equal internal measurement data.

10. Provides a defensible basis for grouping students based on test segments of reasonable lengths (eight to ten items) as well as on basic skills subject measures. Cross grade grouping is based on scaled achievement regardless of grade attending.

Individualization. It is important to understand how it is possible to give tests at different levels of difficulty (normally five levels in a regular classroom) and be able to translate the raw scores on the different tests to the same curriculum scale. What is necessary in order for 80 percent of the items correct on a lower level test, to yield a lower measurement than only 60 percent correct on a higher level test? It is necessary to have a way of giving credit based on the levels of the individual items in each of the tests and that they be synchronized with the levels of items in all the other tests. This is justifiable only if item calibrations have been empirically determined and the stability of the calibrations known to be satisfactory as was shown for the Northwest Evaluation Association Banks in the tables in Appendix #1. Ingebo (1979), Forster (1981). When we stay with reality, as opposed to statistics, we once again find the psyche and the psychometrics in agreement. The three extremely important conditions for obtaining accurate measurement are: 1) student understanding of the questions, 2) appropriateness of methodology for communicating responses, and 3) test taking mental set of the student. All influence the likelihood that the test score represents the student's capabilities. All or especially the mental set

of the individual can depress or increase test scores and consequently have an adverse affect on accuracy.

The queen of district testing programs is the survey test. It also serves as a good example in showing the importance of knowing about test items. Tests are administered to someone who has learned something to some extent. Measurement then needs to focus on the extent to which someone (not just anyone) as learned. The range of student abilities in grade six, for example, extends up and down the scale to overlap the means for all grades three through nine. Since only one standard deviation of these students achievement each way will encompass means for grades four and eight, it is obviously incongruous to try to measure all students in a grade with the same set of items with a single long range test. (It would be equally incongruous to test a high jumper by having him attempt to jump a bar at all levels from three feet up to eight feet.) Although in the past group tests were understood to be best used for assessing groups, survey measurement is not really something one does to a group. Always the individual is measured, just as it is the individual who learns. Obviously an acceptable range of item levels should be selected for measuring achievement of any one student at whatever his level of ability, even if assessment of a group is wanted. However, this added precision in measurement can be attained only if the item levels have been carefully calibrated and are clearly indexed by goal referencing.

When a question beyond the student's knowledge level is misunderstood, the student will sometimes select the right answer for the wrong reason. A careless response to a seemingly obvious easy item can likewise result in a mistake. An angry, frustrated, or defeated mental set can result in successful cheating behavior or partial withdrawal from the situation. In the past some students were required to sit still and take a pencil and paper test including items far beyond their academic achievement level or comprehension. That is mental abuse as severe and possibly as prevalent in our system as physical abuse.

Teachers do a good job of defusing some of the fears of traditional tests. After all, a part of teaching is to encourage students to take on assignments that are difficult and to do their best. However, a few students (one is too many) taking a traditional standardized test will be faced with a long series of items and be able to manage predictable answers to only a few. As the series of questions lengthens with one confusing or mysterious item after another, realizations grow that all is not well. Over time there is also the memory of preceding testing periods with confusion and failures that tends to establish the mental set of low achieving examinees.

Teachers are not in control when a single test is given to all students in their classes - no way to make allowances for one student's academic interests or another's incapacities. Neither can they down play the importance of tests in order to make the defeats they know some of the students face less traumatic. The crowning insult to their sensitivity for low achieving students is use of a test that is written for more than one grade.

There are laws concerning child abuse (physical) and school policies to support them, not because it is more unjust than mental abuse, but because it is noticeable and objectively observable results exist in many instances, with results persisting long enough to be made apparent to others. Results of child abuse by inappropriate testing tends to inflict covert kinds of distress that are fleetingly evidenced by observable behavior -- transitory evidence that is often overlooked because there seems to be no remedy.

We don't need laws in solving this measurement problem. It's in our own front yard - educators are experts in working with the psyche of students. The technology to really alleviate this abusive condition exists for every district and produces not only a more task fitting, individualized approach to testing, but also higher quality information. Psyche fit with psychometry.

To understand the test building power of using an item bank think of it as a long three thousand item test with all items related to each other on an equal interval scale. No students will be given this long range test but short sections of it that include only items appropriate for determining the achievement levels of a segment of the student population with similar performance levels will be used in each levels test. A set of five or so shorter range tests of this kind have been found adequate to accommodate the range of students abilities in most classes at a single grade, third through eighth.

No one likes to be treated unfairly. Even the low achieving students taking appropriate levels tests designed to match the local district curriculum, realize they are being treated fairly. Teachers know that the test results are more accurate, they feel better about the more humane way of evaluating individuals, and they appreciate the easier test administration when all students work all period.

A survey of teachers' opinions about students' reactions was taken the first time the levels tests were administered. Ingebo (1977). Students accepted the fact that different class members would get tests at different levels to match their academic progress with neutral or positive reactions. Teachers volunteered the information that the testing atmosphere was improved by more able students working the full time.

The levels tests system (individualized testing at each grade 3 through 8) comprised the main information gaining component of a complete testing system. No other basic skills tests need to be added for evaluating funded projects, for selecting students to form cross grade groups, for competency standards, etc. However, the same test development principals obtain in building specialized instruments for other than survey measurement purposes. For example, if you want a criterion referenced test with a passing RIT (Rasch Unit) of 200, you select appropriate content items with a narrow range of RIT calibrations ranging, for example, from 198 to 202 inclusive. A levels test for survey purposes with the same center point would most likely have items ranging from 190 to 210 evenly spaced as to their difficulties.

Modifying test instruments as needed is possible because of the way we can build tests from the type item bank -- that is item by item. If you decide to purchase published levels tests (Tests of Individual Progress) to get your system going you can later exchange items for some that better match your curriculum. You can make the tests for your district longer, shorter, easier, harder, or with different curriculum components. You can delete items or not include them in machine scoring -- all without upsetting the longitudinal benefits of consistent measurement. This flexibility is not to be used capriciously. Thoughtfully applied it can mean changed instruments without changed measurement.

The freedom to use individual items as measurement entities is a reward for locating each item on the curriculum continuum. Any group of appropriately administered items from the item bank carries the information needed (Rasch calibrations for each item) to determine where on the curriculum scale any number of a certain group of questions marked correctly (the raw score) will be placed (the conversion table) even before that group of items is administered to students. Once the techniques for test construction are in place, the test modification as well as test development begins and continues with thoughtful and concerned deliberations of involved professionals to determine what information is needed, what test items must ask the students to do, as well as what's in it for students and teachers that will make them take the tests seriously.

Security is important - lack of it can be devastating to individual measurements, misleading in group averages for program evaluations, and embarrassing in board reports. Recently it has became the practice in some districts to require passing a competency test in order to earn a graduation diploma. Security cannot be maintained under this kind of pressure if the same test is used each testing period. The obvious answer is to format a new test each time. However, that can be done with assurance that the same standard is maintained in the new tests only by using an item bank with proven calibrations for the items.

Classroom test security has always been a problem of test administration that varies for groups by protection that ranges from excellent to none, but the fact that students know there are five or six different levels of tests used at the same testing period discourages attempts at copying. Also important is decreasing emotional stress on the lower achievers to reduce tendency to cheat. And students knowing others have little chance to copy will not feel compelled to cheat in self-defense.

It's not likely that all students in a class will look forward to testing day with joyous anticipation but lowering the frustration index for low students who have been encouraged by the teacher to do their best by giving them questions within reason, helps the general feeling tone for administration. Adding the knowledge that other students are not able to make testing a shared experience does away with defensive cheating by middle and upper level students.

Curriculum alignment means current face validity for curriculum and instruction in the district. It is important in some cases that an item or items be written to cover some specific aspect of the curriculum that is not addressed in a levels test or an item bank. Teachers and curriculum personnel will always have this responsibility. In this case, the fixed parameter model programmed by Fred Forester and since used in hundreds of field tests in the Portland, Oregon metropolitan area makes it practical and manageable to test and if they are suitable items to include from one up to a reasonably large number of field test items with accurate calibrations matching those in the N.W.E.A. bank items can be included in a district's expanded item bank. Houser, Hathaway, and Ingebo (1983). This would require using a test based on a regular N.W.E.A. item bank with district's local questions added to it. The fixed parameter model bypasses the long and more arduous process that must be followed if the Rasch model is used to link groups of 30 or more items to other groups of items. Remember, however, to include only curriculum specific items with satisfactory psychometric qualities and rewrite or reject the others.

Equal interval measures provided through use of calibrated items and Rasch scales are not only more compatible than scores based on publishers populations, they also provide the basis for developing local norms in a statistically more appropriate way than simply using the raw scores. Notice the path we have taken here is direct, not some statistical or pseudo-statistical treatment of population data. To emphasize the importance of equal-internal scales, I'll ask you to follow what is admittedly a less than elegant parable, but one that might help clarify what I have in mind.

In an ancient horse loving country, some citizens were concerned with determining how fast their favorite animals could run, but they had no way of measuring short periods of time. To overcome this problem, horses were group tested by having them run a course of convenient distance in a single race. Since careful records were kept of their finishing order, a retread astronomer, who offered his services as a statistical consultant, was able to develop a procedure for assigning each horse a percentile rank. In his search to improve on this idea, the astronomer, who now was assigned to the Bureau of Standards, suggested that each horse's performance be scored in the following way: first a judge would signal precisely when the winning horse has finished, then other judges would count the additional steps required by each other horse to reach the finish line. With these objective data in hand, it was possible to calculate a mean and median number of steps, and then to generate a standard score, semiinterquartile range deviation, stanine, etc., for each horse.

Whereas this country loved Arabians, in a distant country the favorite was the Clydesdale. Soon there were arguments generated by subjective comparisons between horses, and this resulted in a grand scheme to run joint statistics on a random sample of Arabians and Clydesdales. The planning for this program was proceeding smoothly when a letter was received from the Chamber of Commerce of the Shetland Isles, protesting that their favorites were not

adequately represented in the sampling design. Of course, letters from other horse trading centers were not far behind.

All this nonsense came to an abrupt halt, however, when a clever mechanic invented a stopwatch. Using this simple instrument, he was able to assign a measure of performance to each horse, which did not depend on its breeding or the population of other horses entered in the races in which it had run.

The point of this parable is probably obvious. The design of calibrated tests, represents an important step beyond any previous efforts heretofore. Like the stopwatch, the result of levels tests is an equal interval type of measurement comparing a student with curriculum acquisition rather than the population based, rank order type of scales we have had in use in the past. Ingebo (1980).

Many different uses of population based information are available to the district as shown in the list of comparisons presented later in this paper. It is a comforting start to base them on equal interval measures that are scaled directly to the curriculum instead of scores from some large norming group not representing the District's students.

Equal internal measures can make a variety of local district norms important for providing assessment information within an educational unit. Comparative norming means having a large districtwide population take the three basic skills tests of reading, mathematics, and language usage in a short time frame. When the same students take all three tests, standard scores can be used for individuals or small samples to compare information across subjects because the normative data comes from the same norming population.

Three features of a local Rasch bank based system are important to note: 1) since standard score distributions follow the shape of the distributions of measures and since the standard deviation is a statistic that is best used with equal interval measures, RIT measures are an improvement on traditional scores for producing local derived scores; 2) comparing standard scores across two or three subjects accentuates any error in computing such derived scores; and 3) the degree of test mismatch with local curriculum is increased when a battery of tests from the same publisheris adopted because one of the three tests selected as a best fit instead of having the best possible curriculum/item alignment for all tests. It is likely that tests from two or three different publishers will match the three different curriculums in a district better than those in a single battery. Adoption of a battary is convenient but an unnecessary source of error in derived scores.

With publishers' norms (misnamed national norms) the scores of groups of students who were volunteered by districts as norming samples will very likely provide standard deviations of questionable value because they were: 1) based on tests with items mismatched to their curriculum; and 2) from "extra" tests not part of their regular testing program. Such tests are seen as not important to their educational experiences. Even using the original raw score means and standard deviations of the norming group used by the publisher (the

most defensible use of publishers' data) still leaves one with imprecise information. The extreme folly would be to manipulate publishers' grade equivalents mathematically.

A variety of standardized differences between measures such as between fall and spring or between fall and fall testings are defensible when equal interval measures of students achievement are used. They are extremely helpful in evaluating programs when groups of students of class size or greater are available.

Standardized differences were first referred to as "growth index" Houser, R., Ingebo, G. and Karr (1977) in the extensive list of relationships made available to users of a levels testing system. The computations used and purposes for developing growth indices is included later in this paper. They are numerical differences in standard notation based on a distribution of similarly derived differences. The reason one doesn't do this with conventional test is simply that it is not reasonable to add or subtract ordinal scale data (its not statistically holy either) that does not have a basic relationship with the curriculum being measured. However, using equal interval measures to get differences from the full range of low to high groups gives about as good a chance as life has to offer that relationships based on these indices will be dependable for decision making.

Conversion tables are commonly used in achievement testing to translate numbers of items answered correctly to some kind of scale (not including grade equivalent as a scale). A dramatic difference in how this can be done underlines the differing nature as well as improved power of Rasch technology.

It is useful when teachers or people running a program would like a special test for a special purpose and a very small group of students to be able to not only produce the test readily, but along with it a conversion table that will allow the scores on that test to be consistent with scoring in the regular testing program. Using conventional methods it would be necessary to get population based information about these tests and the conversion table could not be used for other than the small group that the test was designed to assess. Notice how distinctly different the conventional population base norming requirements are from the efficient test engineering freedom afforded by Rasch technology.

It is easy, if the general ability of the group is known, to determine how hard to make the special test in order to get the best measurement results. This and the above mentioned capabilities underscore differences in the nature as well as the power of the "Rasch Tiger" in changing to a new kind of testing system.

An overall levels testing system can provide basic skills achievement information for all of the school programs in grades three through eight that are involved in basic skills attainment or improvement. Although investigation of 1st grade measurement was started by Lester Elijah in Tachoma Public Schools, the experimental work hasn't been completed, that we know of,

for group testing with pencil and paper tests in grades one and two or kindergarten. Elijah (1975). But no longer is a variety of redundant testing necessary. Program related group comparisons (based on numbers of students of class size or greater) can also be made with research respectability using routinely gathered test data. See Appendix 2.

Longitudinal information requires a minimum of yearly testing to be cost effective for administrative use. Any central office personnel and the school board can benefit from comparisons across years. At the student and teacher level it is consistent with serious effort such as a year's work for any student interested in learning to be provided objective as well as professional information about progress -- personal for the student, professional for the teacher concerning individuals and for the total class. With the levels tests -- short and individualized -- fall and spring testing is not as intrusive timewise as it is helpful to the learning and instruction that goes on in the classroom..

Program evaluation projects can be made efficiently because matching control groups can be done very tightly and comparisons of growth made with almost total exclusion of regression to the mean. Examples and explanations given later in this session explain program evaluation techniques used successfully over the past five years.

A district's policy of testing students every other year, or less, is an information avoidance waste of student and teacher time exceeded only by the awesome abandonment of important information when a teacher who has learned about students as people over a school year is then shifted to a group of newcomers.

Grouping students using RIT scores provides for an effective methodology Eddy and Olson (1980) because it is not grade related. If a school wants to do cross grade grouping, a student will be placed in a class because of a specific RIT score rather than because of a grade level norm that doesn't necessarily correspond with the norms for the grade levels below or above. Making cut points for purposes like screening for Chapter I classes at various grade levels is facilitated by the RIT scale and the established standard can be maintained even when tests are altered or replaced.

In winding up a review of last years' N.A.P.D. symposium, Dr. Coffman made a couple of statements right on target for anyone using the Rasch; ". . .you have to understand the model, not simply let the computer produce numbers to be interpreted automatically" and; "Finally, if you propose to make use of the Rasch model in generating test scores, recognize that the better you understand the model and its implications, the more likely you are to make sensible interpretations of the scores you generate" are words of wisdom. There is no magic in any mathematics formula, but critical review and hard evidence should be required in guiding the transition to better measurement. To follow-up, recognizing Dr. Coffman's timely cautions, you can make a trial run with your own teachers and students by using already packaged tests and available microcomputer software.

As suggested at the beginning of this picture of Rasch Testing, you can use the flexibility provided to build a strong basic skills testing program -- and enjoy the results. How? Cherish excellence rather than convenience. You can get all the help you need to proceed at a pace reasonable for your district or state. It is professionally reasonable to seek procedures that will bring a program like achievement testing in line with your own value system and how it impacts students and teachers.

LONGITUDINAL CALIBORATION ANALYSIS

When a large group of students from one grade but attending 50 different schools throughout the district take the same levels test of 30 or more items, their individual response records provide substantial data with which to compute Rasch calibrations for the items. If another large group from a different grade takes the same levels test (levels tests are matched to students achievement recrods regardless of grade) and the same items again Rasch calibrated, it is possible to examine inter-item calibration relationships across grades. This is easy because for each Rasch calibration the Rits are centered around 200 (or 0 in logits).

The Calibration Analyses Tables presented in this appendix are copies of worksheet programmed by Fred Forster of the Portland Public Schools Evaluation Department and Curriculum Department could routinely monitor calibration stability of items twice each year.

Although standard errors are provided for the five or six different calibration values of each item, as well as other indications of differences in calibrations, it should be emphasized that like item plots for the item characteristic curves, the most important information comes from studying "what really happened" when students interacted with items (the calibrations for each item at different grades).

CALIBRATION ANALYSIS FOR TEST LO2RD 03/03/82

ITEM	BANKID	03	04	05	06	07	08	09	BANK AVG	MAX DIF	AVG DIFF	STD ERR	MAX JMP	DIFFERENCES BETWEEN GRADES 3 TO 4	4 TO 5	5 TO 6	6 TO 7	7 TO 8	8 TO 9
2	R29201	187	188	187	189	189	188	0	187.8	2.9	0.8	0.4	2.2	1.3	-1.4	2.2	0.2	-1.5	0.
									188	DIFF/TOT				-1.1	0.2	-1.1	1.1	1.2	-0.2	0.
3	R 760	194	194	193	195	197	197	0	194.2	4.4	1.2	1.0	2.3	0.5	-1.1	1.6	2.3	-0.0	0.
									194	DIFF/TOT				-0.7	-0.2	-1.3	0.3	2.5	2.5	0.
4	R29389	188	187	189	189	190	191	0	188.3	5.2	1.3	1.0	2.3	-1.4	2.3	-0.3	1.1	1.7	0.
									189	DIFF/TOT				-0.1	-1.5	0.8	0.4	1.5	3.2	0.
5	R 3540	190	191	187	186	190	191	0	189.0	5.4	1.9	0.6	4.4	0.4	-3.8	-0.8	4.4	0.5	0.
									190	DIFF/TOT				1.2	1.6	-2.2	-3.0	1.4	1.8	0.
6	R29207	183	188	189	188	189	193	0	187.8	9.7	1.8	2.0	4.2	4.2	1.0	-0.8	0.7	4.1	0.
									188	DIFF/TOT				-4.4	-0.2	0.9	0.0	0.7	4.8	0.
7	R29209	188	191	190	192	189	194	0	190.7	5.9	1.6	0.8	4.5	2.9	-1.5	2.2	-2.8	4.5	0.
									191	DIFF/TOT				-2.4	0.5	-0.9	1.2	-1.6	3.0	0.
8	R29430	188	189	188	188	189	192	0	189.0	4.8	1.1	0.9	2.7	1.7	-1.0	-0.5	1.5	2.7	0.
									189	DIFF/TOT				-1.2	0.5	-0.5	-1.1	0.4	3.1	0.
9	R 7320	194	193	191	190	191	191	0	192.0	4.6	1.3	0.6	1.1	-0.5	-1.8	-1.8	1.1	0.7	0.
									191	DIFF/TOT				1.8	1.2	-0.6	-2.3	-1.3	-0.6	0.
10	R29164	202	199	198	197	199	195	0	198.4	7.2	1.7	1.3	2.2	-2.7	-1.2	-1.3	2.2	-3.7	0.
									198	DIFF/TOT				3.3	0.6	-0.5	-1.8	0.3	-3.4	0.
11	R 9280	185	191	191	189	190	193	0	189.9	9.2	2.0	1.8	6.6	6.6	-0.4	-1.8	0.7	3.6	0.
									190	DIFF/TOT				-5.4	1.2	0.8	-1.0	-0.3	3.2	0.
12	R29416	197	198	195	197	196	197	0	196.8	2.7	0.6	0.5	1.5	0.7	-2.2	1.5	-0.8	1.3	0.
									197	DIFF/TOT				0.1	0.8	-1.4	0.1	-0.7	0.6	0.
13	R28092	200	199	201	200	198	201	0	199.7	3.1	0.9	0.4	2.6	-1.2	1.7	-0.7	-2.0	2.6	0.
									200	DIFF/TOT				0.5	-0.7	1.0	0.4	-1.6	1.1	0.
14	R29402	199	197	200	198	199	195	0	198.1	5.6	1.3	1.0	3.1	-2.0	3.1	-2.2	0.6	-3.6	0.
									198	DIFF/TOT				1.0	-1.0	2.0	-0.1	0.5	-3.1	0.
15	R29419	202	199	200	198	196	195	0	198.8	7.3	1.9	1.4	0.5	-3.1	0.5	-1.4	-2.5	-0.4	0.
									199	DIFF/TOT				3.3	3.2	0.7	-0.6	-3.1	-3.5	0.
16	R29286	189	192	193	192	191	197	0	192.3	8.3	1.7	1.6	5.2	3.6	1.1	-1.5	-0.6	5.2	0.
									192	DIFF/TOT				-3.6	3.1	1.1	-0.3	-0.9	4.2	0.
17	R28074	200	202	200	200	199	201	0	200.7	3.0	0.8	0.3	1.9	1.6	-1.9	-0.0	-0.5	1.9	0.
									200	DIFF/TOT				-0.4	1.1	-0.8	-0.8	-1.3	0.6	0.
18	N 3480	202	199	200	197	200	200	0	199.4	5.1	1.1	0.9	2.8	-2.4	0.8	-3.0	2.8	0.6	0.
									200	DIFF/TOT				2.2	-3.2	0.6	-2.4	0.4	1.0	0.

CALIBRATION ANALYSIS FOR TEST LO2RD 03/03/82

ITEM	BANKID	03	04	05	06	07	08	09	BANK	AVG	MAX DIF	AVG DIFF	STD ERR	MAX JMP	DIFFERENCES BETWEEN GRADES 3 TO 4	4 TO 5	5 TO 6	6 TO 7	7 TO 8	8 TO 9
19	R28039	206	204	206	207	205	210	0		205.7	6.0	1.5	1.2	4.8	-2.2	1.8	1.5	-2.7	4.8	0.
									206	DIFF/TOT					0.7	-1.6	0.3	1.8	-0.9	3.9	0.
20	R 8820	199	200	200	198	200	199	0		199.4	2.3	0.5	0.4	1.8	0.1	0.3	-1.6	1.8	-1.2	0.
									200	DIFF/TOT					0.1	0.2	0.4	-1.2	0.7	-0.6	0.
21	R 8800	191	194	194	195	194	194	0		193.9	3.9	0.7	0.8	2.5	2.5	0.2	0.7	-1.0	0.4	0.
									194	DIFF/TOT					-2.4	0.1	0.3	1.0	-0.0	0.4	0.
22	R 8780	193	195	197	197	194	191	0		195.3	6.3	1.9	1.1	2.0	2.0	2.0	-0.4	-3.3	-2.1	0.
									194	DIFF/TOT					-2.1	-0.1	1.9	1.5	-1.8	-3.9	0.
23	R 9560	205	203	203	203	205	205	0		203.4	2.6	0.8	0.5	1.3	-1.9	0.8	-0.2	1.3	0.2	0.
									204	DIFF/TOT					1.1	-0.7	0.0	-0.2	1.1	1.3	0.
24	R 9540	204	205	206	208	205	205	0		205.6	4.9	1.2	0.8	2.0	0.7	1.6	2.0	-3.0	-0.4	0.
									205	DIFF/TOT					-1.7	-3.9	0.7	2.7	-0.3	-0.7	0.
25	R29008	200	202	202	204	200	197	0		201.5	7.1	1.8	1.3	1.7	1.7	0.9	1.1	-3.6	-2.9	0.
									201	DIFF/TOT					-1.6	0.1	1.0	2.1	-1.6	-4.5	0.
26	R 3280	205	205	205	205	209	210	0		205.5	5.6	1.8	1.6	4.5	-0.4	0.2	0.	4.5	0.4	0.
									205	DIFF/TOT					-0.5	-0.8	-0.7	-0.7	3.9	4.3	0.
27	R 4560	195	197	197	199	200	191	0		196.9	8.7	2.2	1.8	2.5	2.5	-0.3	2.3	0.5	-8.2	0.
									197	DIFF/TOT					-2.4	0.1	-0.2	2.1	2.7	-5.5	0.
28	R28087	204	203	204	205	202	207	0		203.9	4.7	1.0	0.9	4.2	-0.3	0.8	0.6	-2.3	4.2	0.
									204	DIFF/TOT					-0.3	-0.6	0.3	0.9	-1.5	2.7	0.
29	R29418	199	198	200	201	199	195	0		198.7	5.9	1.4	1.0	2.0	-1.4	2.0	1.0	-1.3	-4.0	0.
									199	DIFF/TOT					0.3	-1.1	0.9	2.0	0.6	-3.4	0.
30	R51100	205	205	203	203	202	207	0		204.1	5.5	1.5	0.8	5.0	-0.3	-2.0	0.1	-0.6	5.0	0.
									204	DIFF/TOT					1.0	0.7	-1.2	-1.1	-1.7	3.3	0.
31	R29426	199	199	202	201	202	196	0		200.0	6.6	1.7	1.1	2.7	-0.6	2.7	-0.6	1.2	-6.1	0.
									200	DIFF/TOT					-0.5	-1.1	1.6	1.0	2.2	-3.9	0.
32	R29212	207	208	209	210	209	205	0		208.4	4.9	1.2	0.8	1.1	0.6	1.1	0.8	-0.6	-3.8	0.
									209	DIFF/TOT					-1.0	-0.4	0.7	1.4	0.8	-2.9	0.
33	R51054	210	208	208	206	208	208	0		207.9	4.4	0.9	0.8	2.4	-2.0	0.5	-2.5	2.4	-0.6	0.
									208	DIFF/TOT					1.9	-0.0	0.5	-2.0	0.4	-0.2	0.
34	R29394	200	200	198	199	200	200	0		199.4	3.1	0.7	0.5	1.4	-0.4	-2.2	1.4	0.5	0.3	0.
									200	DIFF/TOT					1.0	0.6	-1.6	-0.2	0.3	0.6	0.
35	R29429	200	200	202	202	199	200	0		200.9	3.5	1.0	0.5	1.4	0.4	1.4	0.2	-3.0	1.1	0.
									200	DIFF/TOT					-0.8	-0.4	1.1	1.2	-1.8	-0.7	0.

CALIBRATION ANALYSIS FOR TEST LO2RD 03/03/82

ITEM	BANKID	03	04	05	06	07	08	09	BANK AVG	MAX DIF	AVG DIFF	STD ERR	MAX JMP	DIFFERENCES BETWEEN GRADES 3 TO 4	4 TO 5	5 TO 6	6 TO 7	7 TO 8	8 TO 9
36	RS1356	205	207	204	207	207	205	0	206.1	3.9	1.2	0.5	3.4	1.7	-3.2	3.4	-0.1	-2.0	0.
									206	DIFF/TOT				-0.6	0.9	-2.3	1.1	1.0	-1.0	0.
37	R 7780	211	210	210	209	208	210	0	210.0	3.9	0.9	0.7	1.7	-1.1	0.3	-1.2	-1.4	1.7	0.
									210	DIFF/TOT				1.3	0.2	0.5	-0.7	-2.1	-0.4	0.
38	R29421	214	213	212	212	215	213	0	212.7	3.8	0.9	0.7	3.3	-1.0	-1.1	-0.1	3.3	-2.0	0.
									213	DIFF/TOT				1.1	3.1	-0.9	-1.1	2.2	0.2	0.
39	R29139	207	207	206	208	203	204	0	206.5	5.5	1.5	1.1	2.1	-0.5	-0.6	2.1	-5.0	0.8	0.
									207	DIFF/TOT				0.8	0.3	-0.3	1.8	-3.2	-2.4	0.
40	R 2840	207	208	208	206	206	206	0	207.0	2.7	0.9	0.3	1.2	1.2	-0.2	-1.5	-0.4	0.3	0.
									206	DIFF/TOT				-0.4	0.8	0.5	-1.0	-1.4	-1.1	0.
41	R29130	217	213	213	212	212	214	0	213.3	6.0	1.3	1.3	1.8	-4.2	-0.2	-1.1	0.5	1.8	0.
									213	DIFF/TOT				4.0	-0.2	-0.4	-1.5	-1.0	0.8	0.
42	RS1289	212	209	209	209	209	206	0	209.4	6.8	1.2	1.4	0.3	-2.9	-0.5	0.1	0.3	-3.3	0.
									209	DIFF/TOT				3.0	0.1	-0.5	-0.3	0.0	-3.3	0.
43	R29400	218	216	215	217	214	213	0	216.0	5.6	1.4	0.9	0.8	-2.6	0.6	0.8	-2.5	-1.3	0.
									216	DIFF/TOT				2.3	-0.4	0.2	1.0	-1.5	-2.8	0.
44	R29321	211	210	208	207	210	205	0	208.8	6.1	1.7	0.9	2.9	-1.3	-1.7	-1.3	2.9	-4.2	0.
									209	DIFF/TOT				2.3	1.0	-0.7	-2.0	0.9	-3.3	0.
1	R 9560	190	191	194	191	192	195	0	192.1	5.1	1.5	0.8	3.0	1.3	2.9	-3.0	0.5	3.0	0.
									192	DIFF/TOT				-2.1	-3.8	2.1	-1.0	-0.5	2.5	0.

CALIBRATION ANALYSIS FOR TEST LOZRD 03/03/82

ITEM	BANKID	03	04	05	06	07	08	09	BANK AVG	MAX DIF	AVG DIFF	STD ERR	MAX JMP	DIFFERENCES BETWEEN GRADES 3 TO 4	4 TO 5	5 TO 6	6 TO 7	7 TO 8	8 TO 9
2	R29201	187	188	187	189	189	188	0	187.8	2.9	0.8	0.4	2.2	1.3	-1.4	2.2	0.2	-1.5	0.
									188	DIFF/TOT				-1.1	0.2	-1.1	1.1	1.2	-0.2	0.
3	R 760	194	194	193	195	197	197	0	194.2	4.4	1.2	1.0	2.3	0.5	-1.1	1.6	2.3	-0.0	0.
									194	DIFF/TOT				-0.7	-0.2	-1.3	0.3	2.5	2.5	0.
4	R29389	188	187	189	189	190	191	0	188.3	5.2	1.3	1.0	2.3	-1.4	2.3	-0.3	1.1	1.7	0.
									189	DIFF/TOT				-0.1	-1.5	0.8	0.4	1.5	3.2	0.
5	R 3540	190	191	187	186	190	191	0	189.0	5.4	1.9	0.6	4.4	0.4	-3.8	-0.8	4.4	0.5	0.
									190	DIFF/TOT				1.2	1.6	-2.2	-3.0	1.4	1.8	0.
6	R29207	183	188	189	188	189	193	0	187.8	9.7	1.8	2.0	4.2	4.2	1.0	-0.8	0.7	4.1	0.
									188	DIFF/TOT				-4.4	-0.2	0.9	0.0	0.7	4.8	0.
7	R29209	188	191	190	192	189	194	0	190.7	5.9	1.6	0.8	4.5	2.9	-1.5	2.2	-2.8	4.5	0.
									191	DIFF/TOT				-2.4	0.5	-0.9	1.2	-1.6	3.0	0.
8	R29430	188	189	188	188	189	192	0	189.0	4.8	1.1	0.9	2.7	1.7	-1.0	-0.5	1.5	2.7	0.
									189	DIFF/TOT				-1.2	0.5	-0.5	-1.1	0.4	3.1	0.
9	R 7320	194	193	191	190	191	191	0	192.0	4.6	1.3	0.6	1.1	-0.6	-1.8	-1.8	1.1	0.7	0.
									191	DIFF/TOT				1.8	1.2	-0.6	-2.3	-1.3	-0.6	0.
10	R29164	202	199	198	197	199	195	0	198.4	7.2	1.7	1.3	2.2	-2.7	-1.2	-1.3	2.2	-3.7	0.
									198	DIFF/TOT				3.3	0.6	-0.5	-1.8	0.3	-3.4	0.
11	R 9280	185	191	191	189	190	193	0	189.9	9.2	2.0	1.8	6.6	6.6	-0.4	-1.8	0.7	3.6	0.
									190	DIFF/TOT				-5.4	1.2	0.8	-1.0	-0.3	3.2	0.
12	R29416	197	198	195	197	196	197	0	196.8	2.7	0.6	0.5	1.5	0.7	-2.2	1.5	-0.8	1.3	0.
									197	DIFF/TOT				0.1	0.8	-1.4	0.1	-0.7	0.6	0.
13	R28092	200	199	201	203	198	201	0	199.7	3.1	0.9	0.4	2.6	-1.2	1.7	-0.7	-2.0	2.6	0.
									200	DIFF/TOT				0.5	-0.7	1.0	0.4	-1.6	1.1	0.
14	R29402	199	197	200	198	199	195	0	198.1	5.6	1.3	1.0	3.1	-2.0	3.1	-2.2	0.6	-3.6	0.
									198	DIFF/TOT				1.0	-1.0	2.0	-0.1	0.5	-3.1	0.
15	R29419	202	199	200	198	196	195	0	198.8	7.3	1.9	1.4	0.5	-3.1	0.5	-1.4	-2.5	-0.4	0.
									199	DIFF/TOT				3.3	3.2	0.7	-0.6	-3.1	-3.5	0.
16	R29286	189	192	193	192	191	197	0	192.3	8.3	1.7	1.6	5.2	3.6	1.1	-1.5	-0.6	5.2	0.
									192	DIFF/TOT				-3.6	3.1	1.1	-0.3	-0.9	4.2	0.
17	R28074	200	202	200	200	199	201	0	200.7	3.0	0.8	0.3	1.9	1.6	-1.9	-0.0	-0.5	1.9	0.
									200	DIFF/TOT				-0.4	1.1	-0.8	-0.8	-1.3	0.6	0.
18	W 3480	202	199	200	197	200	200	0	199.4	5.1	1.1	0.9	2.8	-2.4	0.8	-3.0	2.8	0.6	0.
									200	DIFF/TOT				2.2	-3.2	0.6	-2.4	0.4	1.0	0.

CALIBRATION ANALYSIS FOR TEST LO2RD 03/03/82

ITEM	BANKID	03	04	05	06	07	08	09	BANK	AVG	MAX DIF	AVG DIFF	STD ERR	MAX JMP	DIFFERENCES BETWEEN GRADES 3 TO 4	4 TO 5	5 TO 6	6 TO 7	7 TO 8	8 TO 9
19	R28039	206	204	206	207	205	210	0		205.7	6.0	1.5	1.2	4.8	-2.2	1.8	1.5	-2.7	4.8	0.
									206	DIFF/TOT					0.7	-1.6	0.3	1.8	-0.9	3.9	0.
20	R 8820	199	200	200	198	200	199	0		199.4	2.3	0.5	0.4	1.8	0.1	0.3	-1.6	1.8	-1.2	0.
									200	DIFF/TOT					0.1	0.2	0.4	-1.2	0.7	-0.6	0.
21	R 8800	191	194	194	195	194	194	0		193.9	3.9	0.7	0.8	2.5	2.5	0.2	0.7	-1.0	0.4	0.
									194	DIFF/TOT					-2.4	0.1	0.3	1.0	-0.0	0.4	0.
22	R 8780	193	195	197	197	194	191	0		195.3	6.3	1.9	1.1	2.0	2.0	2.0	-0.4	-3.3	-2.1	0.
									194	DIFF/TOT					-2.1	-0.1	1.9	1.5	-1.8	-3.9	0.
23	R 9560	205	203	203	203	205	205	0		203.4	2.6	0.8	0.5	1.3	-1.9	0.8	-0.2	1.3	0.2	0.
									204	DIFF/TOT					1.1	-0.7	0.0	-0.2	1.1	1.3	0.
24	R 9540	204	205	206	208	205	205	0		205.6	4.9	1.2	0.8	2.0	0.7	1.6	2.0	-3.0	-0.4	0.
									205	DIFF/TOT					-1.7	-3.9	0.7	2.7	-0.3	-0.7	0.
25	R29008	200	202	202	204	200	197	0		201.5	7.1	1.8	1.3	1.7	1.7	0.9	1.1	-3.6	-2.9	0.
									201	DIFF/TOT					-1.6	0.1	1.0	2.1	-1.6	-4.5	0.
26	R 3280	205	205	205	205	209	210	0		205.5	5.6	1.8	1.6	4.5	-0.4	0.2	0.	4.5	0.4	0.
									205	DIFF/TOT					-0.5	-0.8	-0.7	-0.7	3.9	4.3	0.
27	R 4560	195	197	197	199	200	191	0		196.9	8.7	2.2	1.8	2.5	2.5	-0.3	2.3	0.5	-8.2	0.
									197	DIFF/TOT					-2.4	0.1	-0.2	2.1	2.7	-5.5	0.
28	R28087	204	203	204	205	202	207	0		203.9	4.7	1.0	0.9	4.2	-0.3	0.8	0.6	-2.3	4.2	0.
									204	DIFF/TOT					-0.3	-0.6	0.3	0.9	-1.5	2.7	0.
29	R29418	199	198	200	201	199	195	0		198.7	5.9	1.4	1.0	2.0	-1.4	2.0	1.0	-1.3	-4.0	0.
									199	DIFF/TOT					0.3	-1.1	0.9	2.0	0.6	-3.4	0.
30	R51100	205	205	203	203	202	207	0		204.1	5.5	1.5	0.8	5.0	-0.5	-2.0	0.1	-0.6	5.0	0.
									204	DIFF/TOT					1.0	0.7	-1.2	-1.1	-1.7	3.3	0.
31	R29426	199	199	202	201	202	196	0		200.0	6.6	1.7	1.1	2.7	-0.6	2.7	-0.6	1.2	-6.1	0.
									200	DIFF/TOT					-0.5	-1.1	1.6	1.0	2.2	-3.9	0.
32	R29212	207	208	209	210	209	205	0		208.4	4.9	1.2	0.8	1.1	0.6	1.1	0.8	-0.6	-3.8	0.
									209	DIFF/TOT					-1.0	-0.4	0.7	1.4	0.8	-2.9	0.
33	R51054	210	208	208	206	208	208	0		207.9	4.4	0.9	0.8	2.4	-2.0	0.5	-2.5	2.4	-0.6	0.
									208	DIFF/TOT					1.9	-0.0	0.5	-2.0	0.4	-0.2	0.
34	R29394	200	200	198	199	200	200	0		199.4	3.1	0.7	0.5	1.4	-0.4	-2.2	1.4	0.5	0.3	0.
									200	DIFF/TOT					1.0	0.6	-1.6	-0.2	0.3	0.6	0.
35	R29429	200	200	202	202	199	200	0		200.9	3.5	1.0	0.5	1.4	0.4	1.4	0.2	-3.0	1.1	0.
									200	DIFF/TOT					-0.8	-0.4	1.1	1.2	-1.8	-0.7	0.

CALIBRATION ANALYSIS FOR TEST LO2RD 03/03/82

ITEM	BANKID	03	04	05	06	07	08	09	BANK	AVG	MAX DIF	AVG DIFF	STD ERR	MAX JMP	DIFFERENCES BETWEEN GRADES 3 TO 4	4 TO 5	5 TO 6	6 TO 7	7 TO 8	8 TO 9
36	R51356	205	207	204	207	207	205	0		206.1	3.9	1.2	0.5	3.4	1.7	-3.2	3.4	-0.1	-2.0	0.
										206		DIFF/TOT			-0.6	0.9	-2.3	1.1	1.0	-1.0	0.
37	R 7780	211	210	210	209	208	210	0		210.0	3.9	0.9	0.7	1.7	-1.1	0.3	-1.2	-1.4	1.7	0.
										210		DIFF/TOT			1.3	0.2	0.5	-0.7	-2.1	-0.4	0.
38	R29421	214	213	212	212	215	213	0		212.7	3.8	0.9	0.7	3.3	-1.0	-1.1	-0.1	3.3	-2.0	0.
										213		DIFF/TOT			1.1	3.1	-0.9	-1.1	2.2	0.2	0.
39	R29139	207	207	206	208	203	204	0		206.5	5.5	1.5	1.1	2.1	-0.5	-0.6	2.1	-5.0	0.8	0.
										207		DIFF/TOT			0.8	0.3	-0.3	1.8	-3.2	-2.4	0.
40	R 2840	207	208	208	206	208	206	0		207.0	2.7	0.9	0.3	1.2	1.2	-0.2	-1.5	-0.4	0.3	0.
										208		DIFF/TOT			-0.4	0.8	0.5	-1.0	-1.4	-1.1	0.
41	R29130	217	213	213	212	212	214	0		213.3	6.0	1.3	1.3	1.8	-4.2	-0.2	-1.1	0.5	1.8	0.
										213		DIFF/TOT			4.0	-0.2	-0.4	-1.5	-1.0	0.8	0.
42	R51289	212	209	209	209	209	206	0		209.4	6.8	1.2	1.4	0.3	-2.9	-0.5	0.1	0.3	-3.3	0.
										209		DIFF/TOT			3.0	0.1	-0.5	-0.3	0.0	-3.3	0.
43	R29400	218	216	215	217	214	213	0		216.0	5.6	1.4	0.9	0.8	-2.6	0.6	0.8	-2.5	-1.3	0.
										216		DIFF/TOT			2.3	-3.4	0.2	1.0	-1.5	-2.8	0.
44	R29321	211	210	208	207	210	205	0		208.8	6.1	1.7	0.9	2.9	-1.3	-1.7	-1.3	2.9	-4.2	0.
										209		DIFF/TOT			2.5	1.0	-0.7	-2.0	0.9	-1.3	0.
1	R 9560	190	191	194	191	192	195	0		192.1	5.1	1.5	0.8	1.0	1.3	2.9	-3.0	0.5	3.0	0.
										192		DIFF/TOT			-2.1	-3.8	2.1	-1.0	-0.5	2.5	0.

ACHIEVEMENT COMPARISONS

This paper illustrates the benefits of an individualized testing program based on an equal-interval scale by providing a list of possible comparisons in achievement available through the use of an Achievement Level Testing Program based on the NWEA item banks. The format used in the presentation of these comparisons is the statement of a question followed by one or more explanations of how the data can answer that question. The last section of this paper provides a procedure for planning a set of reports for all user groups.

Professionals with different responsibilities for the operation of a school district can benefit from student achievement information, because it can be presented in different ways to address a variety of questions. However, it is difficult to make a single report clearly present more than a few of the many distinct kinds of information that can be derived from a well thought out testing program. A list of the distinct kinds of reportable information available can help in prioritizing and combining information to be reported for personnel involved in instruction, building administration, central office administration, and other qualified user groups. One way of generating such a list is to think of every use of test information as a comparison put in number form. For example, if a student is found to be 4 feet, 10 inches tall, the results indicate that when the student was asked to stand by a scale measuring distance in feet and inches, the comparison between the student's height and the scale was established at 4 feet, 10 inches. Similarly, the answers a student gives to calibrated questions on a math test can establish a comparison between the student's achievement and the math curriculum attainment scale expressed in Rasch Units (RITs).

Another example of the comparison, this time using two measures, is as follows: When a student was found to be 4 feet, 10 inches tall some time ago and is now 5 feet, 2 inches tall, the comparison between then and now measures is a difference of four inches. This number, 4 inches, relates measures of height of the same student at different times. Similarly, when a student whose Math RIT was 200 in the Fall at grade 5 now measures 205 in the Spring, a comparison between the measures of the same student at two different times is a difference of 5 RITs. When two measures are compared in this way, a third kind of information emerges. In this case, the 5 RIT difference indicates a change over a specific time period for one student and that change can be compared to the average change for other students in the same educational unit such as a school district. By using the population variance as a guide, the change in measures can be thought of as an amount of growth for that specific time period. However, it must be kept in mind that this growth index is no longer a direct comparison with the curriculum, but is a population-based value statistically compared to performance of a specific group of students over the same period of time.

The following list of comparisons includes those that have been provided parents, teachers, administrators, or Board members, in a variety of report formats. It is important to distinguish between the RIT (Rasch Unit) measure which is an equal interval scale and the population based P-score (Portland Standard Score). RITs are more like the feet and inches and are basic measures from which other kinds of scores and statistics are derived, so we will list their uses first. Following will come P-scores, growth indexes, and rate indexes that can be made only by using Levels Tests (available in published form as Tests of Individual Performance).

CURRICULUM-BASED RIT MEASURES

1. How much knowledge have students acquired about the curriculum offered in a subject?

   1.1 RIT measures compare a student's achievement level with the curriculum in a subject through a measurement scale that works like the temperature scales do when Farenheit or Celsius degrees are related to amount of heat.

2. What progress is being made by individual students from test to test such as Fall to Spring or Fall to Fall?

   2.1 RIT measures are on the same curriculum scale so they can be used to compare achievement levels in the same subject from year to year to year from grades 3-9. Growth spurts and plateaus can be detected by keeping longitudinal records for students.

3. How would students from different grades be placed in a cross-grade grouping program?

   3.1 RIT measures accurately compare all students' levels of achievement with all others in a subject regardless of grades they are in and regardless of how high or low their performance might be in the school. Students' knowledge in a curriculum area can be placed in rank order in grades 3-9 in a school. Groups can be established by students' levels of performance at a given time.

4. Does a student's achievement meet a set of standards such as minimum competency requirement or Talented and Gifted placement or identification for Chapter I classes?

   4.1 RIT measures compare directly the student's level of attainment with the level set for a standard of performance in a subject whether it is a RIT level comparison with the curriculum or a population determination such as above or below a certain percentage of a group.

5. Can a student's score be compared with the means of selected groups such as district, school, or participants in a special program?

   5.1 RIT measures are especially appropriate for calculating means because they are on an equal interval scale. Individual RITs can be compared directly with group means because they are on the same scale.

   5.2 RIT measures for individuals can be interpreted using different groups to calculate the groups' standard deviations since both are on the same equal interval scale.

6. Can achievement levels for two or more groups be directly compared for the same subject and testing period?

   6.1 RIT means, standard deviations and group distributions can be compared directly with those of other groups such as district population or schools or specified samples because they are on the same equal interval scale.

   6.2 A number of means can be put in rank order for comparisons.

   6.3 Tests of statistical significance of mean differences are possible because the scale is equal interval -- providing that the design for the test of significance selected is appropriate in other respects.

7. Are students evenly distributed according to achievement within a group or is there a disproportionate number in some segment using quartiles, stanines, high-middle-low, etc., to identify group ranges?

   7.1 RIT measures can be used to define segments of a group and compare numbers or percents of students in each identified group because they are equal interval along the total scale from grades three to nine.

   7.2 Ranges of RIT measures for determining segments can be based on a class-sized group or on the total population of a district or on a sample selected for special program evaluation.

8. Can progress of individual students be shown by difference in RIT scores from one testing period to another?

   8.1 Since RITs are equal interval measures and are referenced directly to the curriculum, differences between two testing periods can be used to follow a student's progress by comparing a difference with next year's difference in grades 3-8.

   8.2 RIT gains can be used to compare a student's progress with the mean and standard deviation of all other students in a district over a specified time such as one year.

   8.3 A pattern of standardized gains over a series of time periods such as Fall to Spring or Fall to Fall, etc. can be plotted to show progress over time, because the scale is equal interval and extends from grades 3-9.

9. Can the mean RIT gain for one group be compared directly with the mean gain of another group or other groups?

   9.1 RIT gains for groups such as school units where sociological factors other than achievement measures are used to define the group are directly comparable with a minimum of regression to the mean because

students are tested at their level regardless of overall school performance.

9.2 The school groups in a district can be rank ordered by RIT gains for the same time period and subject.

10. Can RIT score gains by students at different achievement intervals such as stanines, quartiles, etc. segments be compared to determine which if any segments are contributing most heavily to the total group mean and in which direction?

10.1 The RIT gains for different segments of a distribution are comparable but an important refinement is to identify the segments by measures from an earlier test such as the proceeding year (select with Test A, pre-test with Test B, and post-test with Test C where all tests are appropriate for the student). This pattern of testing is inherent in a Levels Testing Program.

11. Can the relative progress of specific segments in one distribution be compared with similar segments in another group such as showing the progress of low segments is greater in one group but the progress of higher segments is greater in the other group?

11.1 The RIT measure is appropriate and a cursory reading of comparative gains can be useful, but for any appreciable emphasis in decision making the gains should be standardized according to a district population and number of cases in each segment taken into consideration.

POPULATION-BASED STANDARD SCORES

P-scores differ from RIT measures in that they refer to a population norm that is re-established each testing period. P-scores are standard scores converted to a mean of 50 and a deviation of 10. This is done simply to make them easily read without using decimals or minus values. However, it is important to know that these standard scores are based on equal interval measures (RITs) rather than ordinal scores such as numbers of questions correct. P-score scales are differentially normed for the three basic skills test series: Math, Reading and Language Usage. This use of essentially the same population of students for all three subject norms makes it mathematically holy to compare P-scores across subjects or grades or years.

12. Is a student's performance above or below that of the other students in the district? How much?

12.1 A student's P-score is in itself a comparison of a student's level of achievement with the district average in tenths of a standard deviation.

13. Has the student maintained or changed achievement position with respect to other students in the district?

    13.1 The P-score is recomputed for each testing period so it always shows a student's position compared to the district population. A higher or lower P-score in a subsequent testing period simply shows that gain or loss has been greater or less than that of other students in the district.

14. Is a student's position in the district population for one subject, such as arithmetic, higher or lower than the position for reading or language usage?

    14.1 P-scores computed for different subjects using a district population are based on essentially the same students and when the mean and variance needed is based on equal interval RIT scores, the P-score comparisons are defensible.

15. How does a student's achievement compare with that of other specified students?

    15.1 Since P-scores are standard scores they compare directly with other individuals' P-scores in the same subject and for the same testing period.

16. Can a student's achievement be compared with the mean achievement of members of a specific instructional group?

    16.1 P-score means from specified groups other than the total district population on which P-scores are based, such as a school or members of a particular program, can be compared with individual P-scores because the distribution of P-scores has the same shape as the equal interval RIT score distribution from which it is derived.

17. Can an achievement criterion such as that often used for Chapter I selection be established in P-scores?

    17.1 A student's P-score can be compared with an established standard for a given subject.

18. Does a specified group have different achievement means for math, reading, and language usage?

    18.1 Groups other than the district population P-score means in math, reading, and language usage can be compared across subjects for the same testing period. P-scores are derived for each subject using the same mathematical formulas and populations.

19. Does a specified group have different achievement means over successive testing periods in the same subject?

    19.1 Group P-score means in a subject for groups other than the total district can be compared across different testing periods.

20. Are students in a specific group disproportionately represented in different performance ranges?

    20.1 A group can be segmented by P-score ranges (quartiles, stanines, etc.) to compare numbers of students, means, percentages, etc. in each segment.

21. Can the distribution of students in a specific group such as a disproportionately large number of high or low scoring students taking a test be compared with their performance on another basic skills test?

    21.1 Means within P-score ranges for a group can be compared across subjects for the same time period.

22. Is there a difference in the distribution of achievement means over successive testing periods?

    22.1 The means within specific P-score ranges such as stanines, quartiles, etc., for a group can be compared over testing periods for the same subjects.

23. Can the distribution of students in a specific group such as a disproportionately large number of high or low scoring students taking a test be compared with their performance on the same subject at a different testing period?

    23.1 The means within specific P-score ranges for a group can be compared across subjects for the different testing period.

24. Can the distribution of students in a specific group such as a disproportionately large number of high or low scoring students taking a test be compared with their performance on another group's performance in the same subject and testing period?

    24.1 The means within specific P-score ranges in one group can be compared with those of other groups for the same subjects and testing periods.

25. How do two or more groups of students compare in their progress from year to year?

25.1 The relative progress from one testing period to another of two or more groups is indicated by P-scores with respect to the up or down drift of the total district population. Either means or total distributions can be compared.

RELATIONSHIPS USING GROWTH INDEXES BASED ON RIT MEASURES

An index of "growth" over a period of time is much more accurate and informative than the numerical difference obtained by subtracting one group mean from another (called gain). Growth as used here takes into consideration which subject, which time span, and a minimum group size. The notation used is in standard deviations for maximum flexibility of use in making comparisons of groups. This adaptability to a variety of situations makes this index invaluable for evaluating instructional programs.

The growth index is computed from a distribution of the differences between means and is expressed in standard scores. These differences were obtained separately for each grade level for all schools in the district. Groups used were those students from a school who took both pre- and post-tests. This is important because the variance obtained represents a selection of socio-economic rather than academic groupings. Only groups of 20 or more students were included. The mean and standard deviation of this distribution of differences determine the growth index.

The growth index gives meaning to progress in achievement that can be used for group comparisons of progress across grades and subjects much like the P-score makes different testing period measures and different subject measures comparable. Basic to the extensive use of standardization is availability of the individualized, equal interval RIT measures. In a comprehensive testing program, regression from one testing period to another is minimized by selection of the pre-test on one or more previous tests and then using different individualized tests for pre- and post-measures.

26. Is a group's growth for a time period greater or less than that expected of grade level groups in the district? How much greater or less?

    26.1 The growth index is a standardized comparison of a group's growth with that expected of groups in the district and is interpreted as a standard deviation. This index is defensible because its calculations are based on equal interval, RIT measures.

27. In which basic skill subject did a specific group make the most progress?

    27.1 The growth index compares progress in one subject with that in another directly over the same time period. Whatever differences exist can be interpreted as standard score differences.

28. Can a growth index indicate which of two or more groups made more progress? Can grade groupings by school be rank ordered? Can large groups drawn from many schools, such as Chapter I students, be compared with a smaller group from only one school?

28.1 Growth indexes give comparisons in standard notation between groups of 20 or more students. All reasonable comparisons can be made between groups across either grades or subjects because indexes are derived from a nearly total district population which makes an unquestionably representative sample.

29. Can the group from one grade to the next be compared in terms of growth if a large fraction of the students in the first year have been replaced by the second year?

29.1 The growth index is a measure of a specific group having pre and post measures. Some schools need to serve a changing student body and the programs need to be planned and evaluated with that fact considered. If the number of students tested both testing periods is less than twenty other test information than the growth index should be used.

RIT MEASURE BASED RATE INDEXES

"Rate" is the name chosen for an index of change in the growth indexes for a specific group. The growth index for one period of time, 1985 for example, can be compared with the growth for the same group the previous year, 1984, because their means are both in standard score notation. If the later, 1985, growth is larger the group has increased its rate of growth and the rate index will be positive and will indicate magnitude, as well as direction of change in standard deviations.

Much like growth index, rate index is calculated by putting a large number of differences between growth indexes for a grade into a distribution and calculating the mean and standard deviation of these changes. The resulting control of both comparable grouping and regression to the mean is invaluable for program evaluations.

30. Is a group rate of growth different from that expected in a school district? How much:

30.1 The rate index compares the growth rate of a specific group with that of other district groups in standard deviation terms.

31. Is there a difference in the effectiveness of a particular schools' instructional programs across the three subjects?

31.1 Rate indexes can be used to compare the progress of a group of students in two or more subjects because the total district population is involved in providing the data used to compute the three indexes.

32. Can a change in an instructional program or adoption of a new methodology be evaluated districtwide using rate of growth? For a single school?

32.1 Rate has been developed primarily to facilitate evaluation of new or changed methodologies by comparing them with previously used instructional programs. Its main advantages are that it minimizes regression to the mean, meets control group requirements and the data flows conveniently out of an ongoing district testing program. The number of cases used needs to be 20 or more.

33. Can student groups identified by different levels of achievement, such as quartiles or high, average and low groupings, be compared as to rates of progress?

33.1 The rate index minimizes regression to the mean when a group is close to the same comparative level the second testing period as it was the first. The necessary condition is that sub-group segments have 20 or more students for each growth index.

34. Can a district with only a small number of schools establish its own growth and rate indexes?

34.1 Since schools change from year to year, it is reasonable for a small district to collect growth data over a three year period and use 30 or more school means to provide the variance needed to determine growth and rate indexes. However, this will produce indexes that need to be interpreted as standards existing over time rather than reflecting year-to- year district level. These indexes should be important for program evaluation.

COMPOSITE SCORES BASED ON RIT MEASURES

Since Rit scores are based on different curriculum scales for math, reading, and language arts they can not be averaged across subjects. However, the nature of Rit scores is such that standard scores such as P-scores based on them for the same population of students can be mathematically combined. The composite score (C) is not a simple average standard scores, however, the sums of the 3 P-scores for all students in the District are put in a distribution and standard scores calculated for these sums. These composite scores are then converted to a mean of 50 and a standard deviation of 10.

The composite is unlike an average because it takes into consideration the consistency of low or high scores over the three subjects. To demonstrate, compare the average of 3 P-scores over 60 or under 40 with the composite (C-score) and the C-score will always be farther from the mean.

It should be noted that the composite basic skills score has limited use due to the difficulty in specifying just what is being measured. It is something like the Decathalon that combines different skills in track events.

35. A C-score compares a student's overall basic skills position with an average for the District.

35.1 How much is a student's basic skills achievement above or below the District mean?

36. C-scores can be used for student or group comparisons in the manner of P-scores except for cross subject comparisons.

There are three kinds of average group measures over periods of time: whole group, clear group, or clear and intact group (fall-spring, fall-fall, spring-fall). There are three major categories: sex, ethnic, and Title I, that be be identified in most districts and comparisons made according to the list above, but are not included here.

EFFICIENT TEST INFORMATION REPORTING

Development of the different types of reports needed in a school district uses three primary kinds of information: 1) a list of comparisons available; 2) a list of questions/decisions/purposes to be addressed by each different user group; and 3) possibilities and limitations for graphic and textual presentation of information. The goal of report writing is to evolve an articulated menu of test reports that meet user groups' informational needs. Although the process will probably be cyclic and reports will need to be field tested, the development of published reports can follow the following steps:

1. Identify user groups. Students, parents, teachers, counselors, principals, central office administrators, and School Board members are groups made up of individuals with widely varying interests in and understanding of test information. Reports that are on target for help in making decisions can serve the more knowledgeable well and help the uniformed learn to use information a testing program makes available.

2. Determine what evaluative or measurement information each group can make use of in line of professional activities. Determination might be somewhat tentative in some instances but, attractive as it is, erring on the side of over-inclusion of comparisons in a report does present problems for some users reading it for the first time.

3. Match the comparisons available from a Levels Testing System with the information needs of each group. You might want to draw up a matrix for ease in considering an optimum arrangement, number, and priority for developing test reports.

4. List non-measurement information such as title, date, subject etc. for each report.

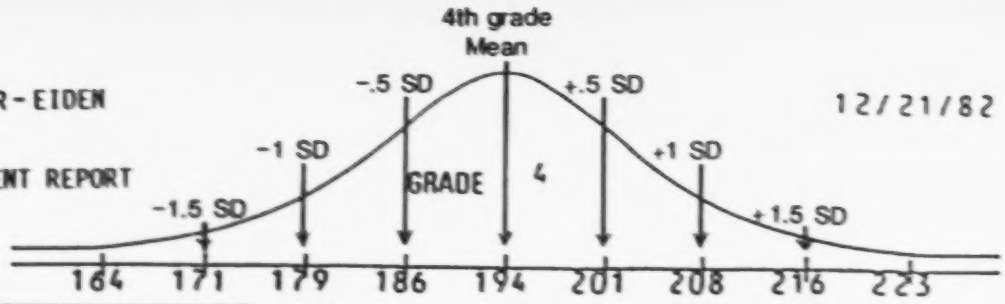
5. Explore data processing resources and programming availability to support quick turn around.

6. Lay out a systematic plan for development tasks and persons responsible for completion of the various reports. Plan for adjustments to be made on the basis of field experience.

1261D

SCH - MEB DEMONSTRATION CTR TCHR - EIDEN 12/21/82

FALL 1982 ACHIEVEMENT LEVEL TEST STUDENT REPORT



RANK MATHEMATICS RIT SCALE: P=40 P=60

STUDENT NAME	P (percent)	RIT
ANDERSON DIANA	62	211
BARTLETT CLIFF	59	207
BONNER KAREN	56	203
CLARK JIM	56	203
COLE JACK	55	201
DOHERTY MAXINE	54	199
DVORAK JAY	52	197
HANSEN LEE	50	194
JENNER JANE	45	186
JOHNSON HENRY	43	184
KIEFERT HARRY	41	181
LARUE JULIE	41	180
LASORDA GLORIA	37	175

VERY LOW | LOW | LOW AVERAGE | AVERAGE | HIGH AVERAGE | HIGH | VERY HIGH

Standard error of 2 Rits

Standard error of 3 Rits

AB = ABSENT, IN = INVALID

LO - THIS SCORE WAS BELOW THE VALID RANGE OF THE TEST

HI - THIS SCORE WAS ABOVE THE VALID RANGE OF THE TEST

Compares student and district average

Compares student with district standard deviation

Compares student with other students in the class

Compares student with math curriculum grades 3 thru 8

Adjective classification

REFERENCES

Elijah, L. (1975). An application of Rasch analysis in the primary reading skills. NorthWest Evaluation Association Journal, 1(1), pp. 7-12.

Forbes, D., & Ingebo, G. S. (1975, April). An empirical test of the content homogeneity assumption involved in Rasch item calibration. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Forster, F. (1981). [Departmental analysis of all levels of tests for mathematics, reading, and language to investigate stability of item calibrations over grades 3 through 8]. Unpublished raw data. Research Department, Portland Public Schools, Portland, OR.

Forster, F., & Ingebo, G. S. (1976, May). Linking groups of items. Paper presented at the spring conference of the Washington Educational Research Association, Seattle, WA.

Houser, R., Hathaway, W., & Ingebo, G. S. (1983, April). An alternate procedure to obtain ability estimates in latent trait models. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Houser, R., Ingebo, G. S., & Karr, C. (1977, November). A new approach to analysis and interpretation of group achievement data. Paper presented at the 56th annual conference of the California Educational Research Association, Los Angeles, CA.

Ingebo, G. S. (1977). [Departmental survey of the impact on students when tests of varying difficulty were given in a class]. Unpublished raw data. Area 3 Evaluation Department, Portland Public Schools, Portland, OR.

Ingebo, G. S. (1979, May). Item calibration and achievement score stability over successive grade levels. Paper presented at the spring conference of the NorthWest Evaluation Association. Seattle, WA.

Ingebo . G. S. (1980). Some perspectives on equal interval curriculum scaling. NorthWest Evaluation Association Journal, 1(1), pp. 30-33.

Olson, A, & Eddy, P. (1980). Rasch unit scores guide the placement of students into mathematics and reading classes. NorthWest Evaluation Association Journal, 1(1), pp. 1-9.

Wolmut, P. F. (1975). A new paradigm for linking goals and items. NorthWest Evaluation Association Journal, 1(2), pp. 7-15.